Imagen que contiene Texto

Descripción generada automáticamente

**MODELO DE VALORACIÓN DE LA VIVIENDA DE COMPRAVENTA**

**GRUPO 4**

**Caso**



04/10/2022

Raúl Carrascosa

Antonio Tomás

Celia Tomás

Verónica Sánchez

**ÍNDICE**

[**1 ABSTRACT**](#_heading=h.2et92p0) **3**

[**2 PROBLEM STATEMENT**](#_heading=h.tyjcwt) **4**

[**3 DESCRIPTIVOS DE DATOS: 8**](#_heading=h.3dy6vkm)

[3.1 Carga de datos](#_heading=h.3znysh7) 8

[3.2 Check](#_heading=h.1t3h5sf) 8

[3.3 Limpieza de datos](#_heading=h.1t3h5sf) 9

[3.4 Transformación 1](#_heading=h.1t3h5sf)1

[3.5 Distribución 1](#_heading=h.1t3h5sf)1

[**4 MODELIZACIÓN**](#_heading=h.4d34og8) **13**

[**5 RESULTADOS**](#_heading=h.2s8eyo1) **y comparativas DE MODELOs 15**

[**6 PRODUCTIVIZACIÓN**](#_heading=h.17dp8vu) **19**

[**8 CONCLUSIONES Y PRÓXIMOS PASOS**](#_heading=h.3rdcrjn) **21**

# abstract

La Ciencia de Datos, lejos de ser una tendencia más, ha supuesto una revolución para todos los sectores e industrias. El sector inmobiliario, en constante cambio debido a las circunstancias socioeconómicas y demográficas, locales y globales, se ha visto beneficiado por la aplicación de la Ciencia de Datos a diversos procesos.

El siguiente trabajo, tiene como objetivo desarrollar un modelo que permita estimar el precio de una vivienda para su compraventa en el portal inmobiliario Idealista para la ciudad de Madrid.

El desarrollo de este modelo se ha realizado mediante iteraciones para aumentar el grado de fiabilidad y disminuir el error de los modelos planteados, realizando un trabajo de limpieza y selección de variables.

Partiendo de un modelo de Regresión Lineal, continuamos con el desarrollo de un modelo de Random Forest para finalizar con un modelo Gradient Boosting. Como variable objetivo, utilizamos precio/m2 y como variables respuesta: la distancia al centro, la distancia a la Castellana, el número de habitaciones, el número de baños, si la vivienda tiene ascensor y el número de metros construidos.

Tras analizar los resultados de las diferentes iteraciones, seleccionamos el modelo de Random Forest para solucionar el problema planteado, obteniendo un valor de r2 de 0,7802 en el set de test y un 0,8055 en el set de entrenamiento.

A continuación de este paso, se presenta una propuesta de productivización mediante la que trasladar los resultados al negocio y poder poner en uso el modelo desarrollado. Proponemos la introducción, a modo de mejora, de una funcionalidad que permita al cliente obtener en tiempo real el resultado de la valoración de la vivienda que quiere vender o comprar. Dando respuesta al objetivo de este trabajo y abordando la necesidad secundaria de maximizar la experiencia del usuario, lo que permitiría a la empresa posicionarse en el sector y distinguirse de la competencia.

# problem statement

Andrew NG, experto reconocido en el campo de la IA, escribió un [artículo](https://medium.com/@andrewng/opening-a-new-chapter-of-my-work-in-ai-c6a4d1595d7b#.y5pmi1j6v) en 2017 cuando se encontraba en momento de salida de la compañía donde lideraba el Grupo de Inteligencia Artificial desde 2014. En el texto mencionado cita: “Así como la electricidad transformó muchas industrias hace aproximadamente 100 años, la IA ahora también cambiará casi todas las industrias importantes y enriquecerá las vidas de innumerables personas”.

Y es que, como vaticinaba el experto, en 2022 no hay sectores que queden exentos de la revolución que ha supuesto la introducción de la Inteligencia Artificial y de la Ciencia de Datos.

En el [artículo](https://www.engelvoelkers.com/es/blog/perspectivas-inmobiliarias/tendencias-del-mercado/inteligencia-artificial-en-el-sector-inmobiliario/) de la empresa de transacciones alemana Engel & Völkers se afirma que la introducción de la Inteligencia Artificial (IA) y la Ciencia de Datos tiene el potencial de revolucionar e innovar el sector inmobiliario, modificando el papel que realizan los actores actuales como vendedores, corredores, gestores de activos e inversores y logrando que los procedimientos se lleven a cabo de forma más eficiente. Traduciéndose además en un potencial de ahorro en los costes de las transacciones inmobiliarias. Sin embargo, igual que en otros artículos consultados, se menciona que estas herramientas virtuales no deben sustituir la importante función de las personas como responsables de la toma de decisiones. En el sector inmobiliario, por lo tanto, se requiere una interacción eficiente entre la inteligencia artificial y el poder de decisión humano.

La empresa de Software inmobiliario Prinex añade, a través de un [artículo](https://prinex.com/el-big-data-en-el-sector-inmobiliario/) consultado, que a través de la introducción de la Ciencia de Datos la industria consigue anticipar tendencias y necesidades, permitiendo una optimización de las inversiones. Conocer qué tipo de inmueble se va a solicitar en el futuro y en qué zonas permitiría a las empresas promotoras y constructoras planificar su oferta de un modo más fiable y con máxima rentabilidad. Además, menciona que el sector inmobiliario se enfrenta a un cliente cada vez más exigente debido a la gran cantidad de información disponible para respaldar la toma de decisiones, siendo la transformación digital el medio para conseguir prestar el servicio que los nuevos clientes demandan.

En el [artículo](https://www.scielo.cl/scielo.php?pid=S0718-33052019000100113&script=sci_arttext&tlng=pt) de Jorge Iván Pérez Rave, investigador del Grupo de investigación IDINNOV, se plantea el desarrollo del modelo *Statihouse,* que caracteriza estadísticamente propiedades en venta y predice su precio de oferta (en tiempo real o cercano).

El mencionado artículo plantea las ventajas que el desarrollo de este tipo de modelos podría tener, siguiendo en línea con la información anterior, para los diferentes actores del sector y para los clientes finales, haciendo hincapié en la mejora de la experiencia de usuario que podría plantear.

En una [Tesis de La Universidad Pontificia Bolivariana](https://repository.upb.edu.co/bitstream/handle/20.500.11912/5285/Modelo%20predicci%C3%B3n%20precios%20viviendas.pdf?sequence=1&isAllowed=y) titulada “Modelo de predicción de precios de viviendas en el municipio de Rionegro para apoyar la toma de decisiones de compra y venta de propiedad raíz” se recalca de nuevo la necesidad de elaborar modelos que predigan el precio de la vivienda. Apuntando que en un contexto industrializado y cada vez más globalizado, donde los precios de los bienes y servicios están en constante cambio, entre ellos el de las viviendas, el cual ha venido aumentando en las principales ciudades, tener una idea de cuánto puede incrementar un precio de una vivienda en una determinada ciudad o municipio puede ser de gran ayuda, tanto para las personas que desean adquirirla para habitar en ella como para quienes la quieran como opción de inversión y así se pueda apoyar el proceso de toma de decisiones en el sector de la vivienda.

En concordancia con lo anterior, también se señala que el hecho de que el valor de mercado esté incrementando podría generar un fenómeno especulativo, donde un inversor no disponga de las herramientas necesarias para conocer el precio real de una vivienda de acuerdo con las condiciones del mercado del sector en el cuál desea invertir/vender.

La Tesis citada también indica que el valor catastral (el cual se obtiene mediante la investigación y análisis estadístico del mercado inmobiliario) podría tomarse como referencia para determinar el precio real de una vivienda, que según fuentes oficiales, corresponde aproximadamente a un 70% del valor comercial (precio del mercado), sin embargo, en muchas propiedades este porcentaje puede ser mucho menor, ya que en la mayoría de las veces el precio de la vivienda depende de cómo se esté comportando el mercado en un determinado sector y en un determinado tiempo, por tanto el valor del catastro no siempre refleja el verdadero precio de la vivienda.

Al mismo tiempo, se alude a que invertir es una decisión que debe ser fruto de un proceso de meditación por lo que antes de realizar una inversión es importante hacer un elaborado estudio del mercado.

La existencia de modelos de estimación del precio de la vivienda ahorraría, por tanto, procesos que conllevan gran cantidad de tiempo, esfuerzo y dinero para los posibles inversores y/o vendedores, haciendo que se evite la toma de decisiones basadas más en la especulación que en la información.

La primera recomendación de las inmobiliarias al momento de comprar una vivienda es determinar el presupuesto con el que se cuenta para ello, ya que una vez se tenga esto, se buscarán condiciones de las viviendas que se ajusten. La segunda recomendación es tener claro qué condiciones de la vivienda se buscan para empezar a hacer relaciones entre vivienda y precio.

Es por esto que tener claridad sobre cuál es la tendencia de los precios de vivienda en el sector que se desea adquirir (ya que los precios del catastro como hemos visto no siempre están acordes con los valores comerciales de los inmuebles), además de las condiciones de dichas viviendas, ayuda a mejorar la toma de decisiones de inversión, ya que se cuenta con un precio inicial de las viviendas pudiendo negociar de forma justa y equitativa la compra o venta de dichos inmuebles y además permite que las personas decidan si se ajusta a su presupuesto y sus necesidades o es necesario buscar bajo otros criterios. Debido a esto, el tener un modelo o plataforma que permita determinar el precio de la vivienda sería altamente eficaz y beneficioso para el inversor.

Una vez justificada la necesidad de esta clase de modelos de predicción del precio de la vivienda, nos proponemos realizar un análisis de la competencia para observar cómo ofrecen este servicio empresas del sector. El porta[l inmobiliario fotocasa](https://www.fotocasa.es/es/tasacion-online/) ofrece un servicio de tasación online y menciona diferentes variables como claves en el establecimiento del precio de la vivienda, un grupo relacionado con la ubicación del inmueble y otro con las características propias del mismo.

En cuanto a la ubicación mencionan:

* La localización y el número de habitantes.
* Los servicios que hay cerca: colegios, centros de salud u hospitales, farmacias, bibliotecas…
* La buena comunicación y los servicios de transporte (paradas de metro, autobús o tren).
* La actividad de la zona: bares, tiendas, restaurantes, espacios dedicados al ocio (cines, teatros, museos, gimnasios, ludotecas…), si es una zona por donde la gente pasea, etc.

En nuestro trabajo, hemos seleccionado dos variables que podrían agrupar las anteriores: Distancia al Centro y Distancia a la Castellana, entendiendo que estas áreas son las que ofrecen el mayor número de los servicios mencionados.

En cuanto a las características propias de la vivienda, se hace alusión al análisis de sus espacios y estado general de la vivienda y el edificio en el que se ubique.

Como dato relevante, mencionar que el peso de las variables ha sufrido un cambio a raíz del confinamiento ocasionado por la pandemia de COVID-19, como confirman en el portal consultado:

“Que el inmueble se encuentre cerca de algunas zonas ajardinadas revaloriza la vivienda, especialmente ahora que, tras el confinamiento, hemos comenzado a [apreciar los espacios abiertos](https://www.fotocasa.es/fotocasa-life/compraventa/cambios-busqueda-vivienda-durante-coronavirus/). Por supuesto, aquí juegan un factor importante las zonas comunes que se ofrecen en las urbanizaciones: piscina, gimnasio, sauna, jardines, patios… Un plus en cualquier vivienda.”

Estos datos también son confirmados por un [artículo](https://agenciamediterranea.com/cambios-de-habitos-en-la-compra-de-la-vivienda-tras-el-covid/) de la inmobiliaria Agencia Mediterránea que señala que el interés por comprar una nueva casa con terraza o jardín se ha incrementado en un 23% y, que cuenten con zonas verdes en la zona o en la comunidad aumenta su interés en un 13%. Si hablamos de que la vivienda cuente con piscina, entonces este porcentaje puede crecer hasta el 45%.

Sin embargo, dado que los datos utilizados para la elaboración de nuestro modelo son previos a la crisis de la COVID-19, estos factores no serían determinantes, pero sí resultaría interesante contemplarlo de cara a futuros desarrollos.

En el portal inmobiliario consultado como parte del estudio de la competencia, también mencionan otras características relevantes de la edificación como el ascensor, que sea o no una vivienda histórica, [si cuenta con plaza de garaje](https://www.fotocasa.es/fotocasa-life/alquiler/los-espanoles-pagan-de-media-71e-al-mes-por-alquilar-una-plaza-de-garaje/) o trastero, el estado del portal…Respecto a la vivienda, la cualidad más importante a la hora de tasarla que se señalan son las dimensiones y materiales de construcción: “no va a tener el mismo precio un estudio de 30m2 que un ático dúplex con piscina privada”. La distribución de los mismos también influye en función del número de habitaciones en las que se encuentra dividido. Por último, se apunta a otras variables relevantes como el año de construcción, la existencia de reformas y la altura a la que se encuentre situada la vivienda.

La precisión de valoración del precio de la vivienda aportada por este portal es del 85%, el objetivo de nuestro trabajo es realizar un modelo de predicción del precio de la vivienda que se acerque, lo máximo posible a los valores de los modelos actualmente utilizados por empresas del sector, tratando además de aportar mejoras añadidas que permitan a Idealista distinguirse de las empresas competidoras.

# descriptivos de datos

Comenzamos realizando el análisis exploratorio del dataset con el objetivo de entender características principales de las variables, tipos de datos, optimización y visualización de estas.

Llevamos a cabo un proceso standard de análisis exploratorio en el que realizamos los siguientes pasos:

Diagrama

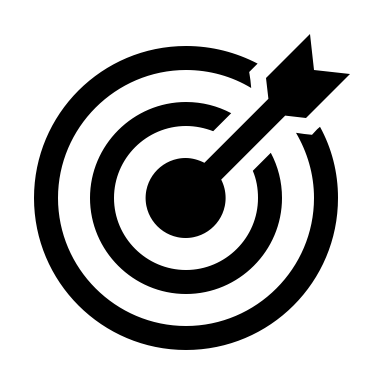
Descripción generada automáticamente

## Carga de datos:

Idealista nos proporciona un dataset de anuncios de venta para la ciudad de Madrid. Consta de **94.727 anuncios** de las que tenemos información a través de **40 variables** entre las que se encuentran: fecha, precio, área construida, tiene terraza, número de baños, ascensor, etc.

## Check:

## Realizamos un describe para ver las primeras características del set de datos y realizamos un Panda Profiling con el objetivo de poder visualizar de una manera más ágil el conjunto de las variables.

Comprobamos que no tenemos valores nulos en ninguna de las variables, que la distribución es multimodal y que el rango de valores a nivel de precio es muy amplio. 

Para esto último realizamos una función densidad de la variable precio con una escala logarítmica. Usamos el logaritmo del precio para equiparar escalas y estandarizarlo.

Interfaz de usuario gráfica, Gráfico

Descripción generada automáticamente

Para poder tener un análisis más exhaustivo por barrios, cargamos los polígonos basándonos en las zonas de Idealista de Madrid.

Calculamos los polígonos censales, el precio por medio cuadrado y así podemos visualizar cómo quedarían por barrios.

Un dibujo de un mapa

Descripción generada automáticamente con confianza bajaDiagrama

Descripción generada automáticamente

## Limpieza de datos:

Con el objetivo de tener unos datos lo más precisos posible llevamos a cabo la detección de valores anómalos o outliers. Para ello realizamos las siguientes acciones:

* Eliminamos anuncios con latitud inferior a 40 🡪 1 registro
* Eliminamos anuncios con área inferior a 30 m2 y superior a 800 m2 🡪 1.601 registros
* Detectamos que existen viviendas que no tienen registrado ningún baño en el anuncio. Creemos que puede ser un error a la hora de introducir los datos y decidimos **eliminarlos** 🡪 83 registros
* Existen anuncios sin habitaciones, los analizamos y vemos que no son estudios por lo que creemos que son lofts con muy poca representación por barrio / zona. Decidimos eliminarlos del modelo ya que no nos aportaran ningún dato relevante 🡪 66 registros
* Nos encontramos con información duplicada en la variable ‘AssetID’, identificador único de cada anuncio, por lo que estamos considerando más registros de los que realmente tenemos.

Lo comentamos con David Rey, CDO Idealista, para conocer la razón por la que se estaba produciendo esta acción. Nos informó que se producía una casuística especial con las inmobiliarias: adquieren paquetes de anuncios y en ocasiones, cuando venden una vivienda, vuelven a reutilizar el mismo espacio en el portal y por lo tanto el mismo AssetID.

Se trata de un **20% de la muestra** duplicada 🡪 75.738 registros de 94.727.

Analizamos más en detalle estos registros para ver si se trataba de estos casos o de una duplicidad del registro en todas sus variables. Para ello comprobamos:

* + Duplicidad en todas las variables anuncio 🡪 18.667 registros
  + Duplicidad en las variables AssetID y área construida: determinamos que son el mismo anuncio 🡪 17.701 registros
  + Duplicidad en la variables AssetID pero con área construida diferente: determinamos que no son el mismo anuncio y se trata de las casuísticas especiales que realizan las inmobiliarias 🡪 966 registros

Tomamos las siguientes decisiones:

* + **Mantenemos** registros con mismo AssetID y diferentes áreas
  + Si tienen mismo AssetID pero diferente área nos **quedamos con el último registro por periodo**
  + **Eliminamos** los anuncios repetidos que tienen el mismo AssetID y misma área
* Calculamos la función de densidad de las variables estadísticas. Primeramente, el valor mínimo y máximo siendo tres veces la desviación típica, para poder detectar valores anómalos y poder descartarlos posteriormente. Vemos que no tienen un comportamiento en referencia a la normal y con valores negativos. Viendo que el volumen es menor al 1% del total de la muestra y que tendríamos que realizar un desplazamiento de la función a la izquierda, decidimos **mantener** estos anuncios 🡪 824 registros Gráfico, Gráfico de líneas

  Descripción generada automáticamente

## Transformación

Vemos que todos los anuncios son de Madrid Capital, por lo que decidimos agrupar los polígonos por las zonas más representativas:Texto

Descripción generada automáticamente

Creamos variables dummies por barrios y comprobamos si hay datos nulos de las posibles variables independientes que podemos usar en el modelo.

Hemos realizado el mismo ejercicio de dummies por zonas con el objetivo de tener más limpieza y exactitud en los datos que usemos en el modelo.

Captura de pantalla de computadora

Descripción generada automáticamente con confianza media

A su vez decidimos rellenar valores nulos con la mediana del total de la muestra.

## Distribución

Obtenemos un mapa de correlación de todas las variables, pero existen tantas variables que el detectar correlación se hace muy complicado. De esta manera decidimos excluir de este estudio / visualización, aquellas variables que no aportan valor al modelo.

Como podemos observar en el mapa de calor, las variables más correlacionadas son las que se acercan a -1 o 1. En el nuestro tenemos bastante relación entre el precio y las distancias al centro / Castellana.

Esto nos indica que puede haber una tendencia de menor distancia mayor precio o viceversa. Nos ocurre lo mismo con el año de construcción, influye de manera significativa con la distancia al precio, por lo que seguramente tendremos las viviendas más antiguas en las zonas céntricas.

Este estudio lo hemos realizado para el total de la muestra, no por zonas ni barrios, por lo que esta correlación entre las variables seguramente sea diferente o más específica para unos casos u otros. Por ejemplo: en las zonas más céntricas seguro que el año de construcción o portero tendrá más correlación que en las afueras.

Gráfico, Escala de tiempo, Gráfico de rectángulos

Descripción generada automáticamente

La correlación examina la relación entre dos variables. Sin embargo, observar que dos variables se mueven conjuntamente no significa necesariamente que una variable sea la causa de la otra. Por eso solemos decir que **"la correlación no implica causalidad".** Realizar este estudio nos aporta detectar qué variables pueden ser susceptibles de excluir en el modelo ya que sus valores nos aportan la misma información que algunas otras variables.

# MODELIZACIÓN

Uno de los objetivos que nos planteamos es realizar un **modelo sencillo end to end,** de cara a tener una visibilidad completa e ir detectando puntos de mejora. De esta forma tendríamos un producto entregable desde casi el principio y así podríamos ir iterando con el objetivo de tener un modelo más robusto, pero siempre con la finalidad de ir viendo resultados en el corto plazo. Realizando pequeñas iteraciones que mejoren los resultados anteriores.

Valoramos realizar varias técnicas de modelaje con el objetivo de determinar un

Como primer modelo decidimos realizar una **Regresión Lineal**. Esta técnica de modelado se emplea para describir una variable de respuesta continua como una función de una o varias variables predictoras, es decir, trata de explicar la relación que existe entre la variable respuesta “ y ” una única variable explicativa “ x “.

Posteriormente, realizamos un modelo **Random Forest**: es un tipo de modelo construido a partir de una agregación de tipo bagging. Los árboles se adaptan bien a comportamientos no lineales, así como a la interacción entre varias variables, aunque, se adaptan tan bien que corren riesgo de sobreajustar (overfitting). Una de las ventajas de este modelo es que los datos no necesitan mucha normalización, solo limpieza de nulos y dumificación de las variables categóricas.

Por último, realizamos un **Gradient Boosting**: es un modelo basado en el boosting de árboles de decisión basado en la combinación de modelos predictivos débiles para crear un modelo predictivo fuerte. La generación de los árboles de decisión débiles se realiza de forma secuencial, creándose cada árbol de forma que corrija los errores del árbol anterior.

Uno de los parámetros de este tipo de argumentos es el learning rate, que controla el grado de mejora de un árbol respecto del anterior.

Un dibujo de un árbol

Descripción generada automáticamente con confianza baja

La diferencia principal entre estos modelos es que mientras en el Random Forest los árboles son formados por conjuntos de variables independientes, el Gradient Boosting construye árboles de manera secuencial.

Definimos como variables respuesta:

* Distancia al centro
* Distancia a la Castellana
* Número de habitaciones
* Número de baños
* Si la vivienda tiene ascensor
* Metros construidos

Como variable objetivo utilizamos el logaritmo del precio ya que estandariza y equipara las escalas para todos los valores. Más adelante, cambiaremos la variable por “precio / m2 ” ya que de esta manera podremos comparar viviendas con diferentes niveles socioeconómicos.

# Pre Modeling

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import mean\_absolute\_error

from sklearn.metrics import mean\_squared\_error

from sklearn.metrics import r2\_score

from sklearn import linear\_model

import seaborn as sns

# create X and Y

feature\_cols = ['BATHNUMBER','HASLIFT', 'HASAIRCONDITIONING', 'HASDOORMAN',

'CADCONSTRUCTIONYEAR', 'FLOORCLEAN', 'FLATLOCATIONID','CADMAXBUILDINGFLOOR',

'CADASTRALQUALITYID', 'DISTANCE\_TO\_CITY\_CENTER', 'DISTANCE\_TO\_METRO', 'DISTANCE\_TO\_CASTELLANA',

'CUSEC']+[col for col in gdf\_join\_out if col.startswith('LOCATIONNAME\_')]+[col for col in gdf\_join\_out if col.startswith('ZONA\_')]

X = gdf\_join\_out[feature\_cols]

y = gdf\_join\_out['UNITPRICE']

y.shape,X.shape

X.columns

# Definimos el modelo

modelo\_linear = LinearRegression()

# Entrenamos y evaluamos el árbol

entrenar\_modelo\_y\_predecir(modelo\_linear)

# Definimos el modelo de Random Forest

modelo\_RF = ensemble.RandomForestRegressor(

n\_estimators = 200,

max\_features = "auto",

max\_depth = 10,

min\_samples\_split = 10,

min\_samples\_leaf = 5)

# Entrenamos y evalúamos el modelo

entrenar\_modelo\_y\_predecir(modelo\_RF)

# Definimos el modelo de Gradient Boosting

modelo\_GB = ensemble.GradientBoostingRegressor(

n\_estimators = 1000,

learning\_rate = 0.01,

max\_depth = 3,

min\_samples\_split = 10,

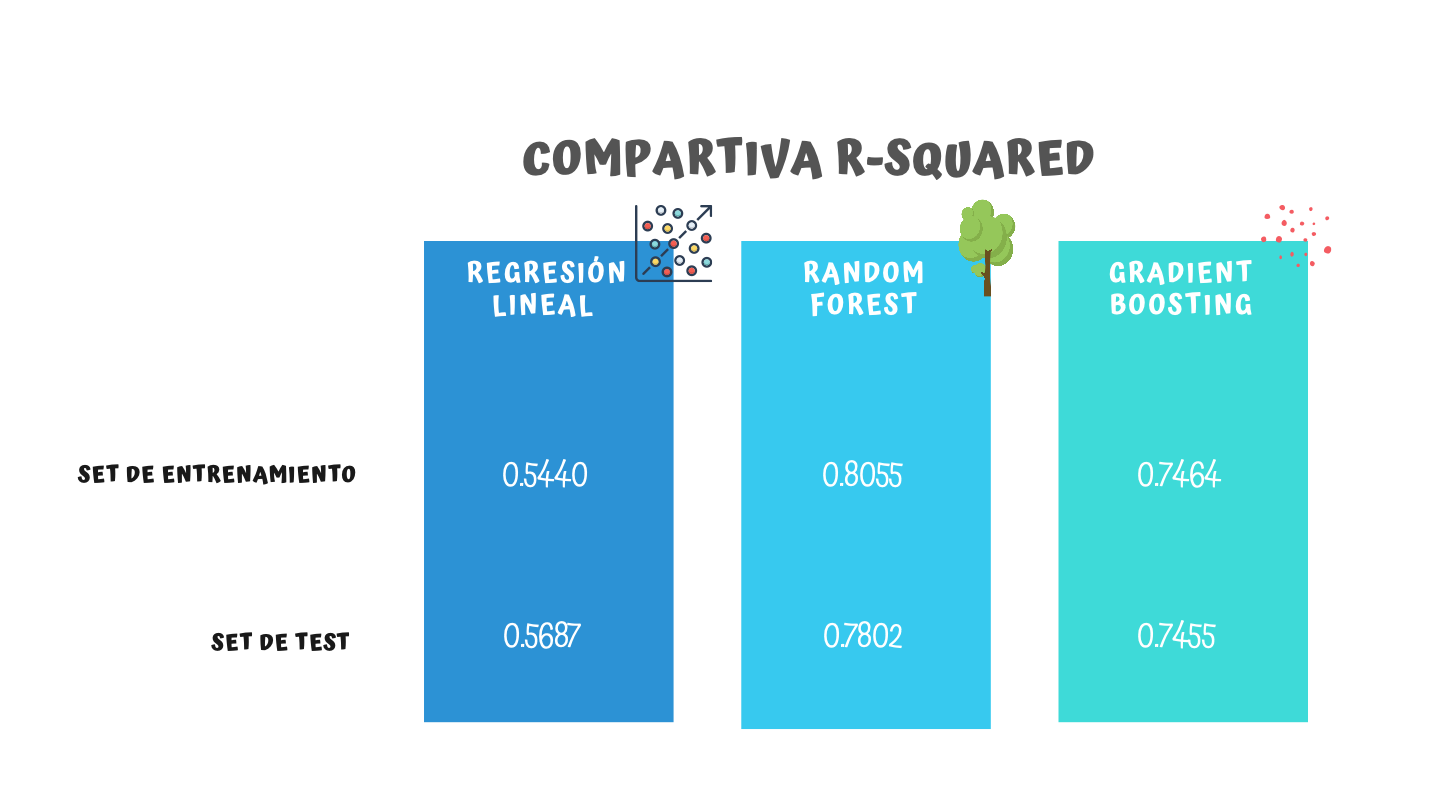
min\_samples\_leaf = 5)

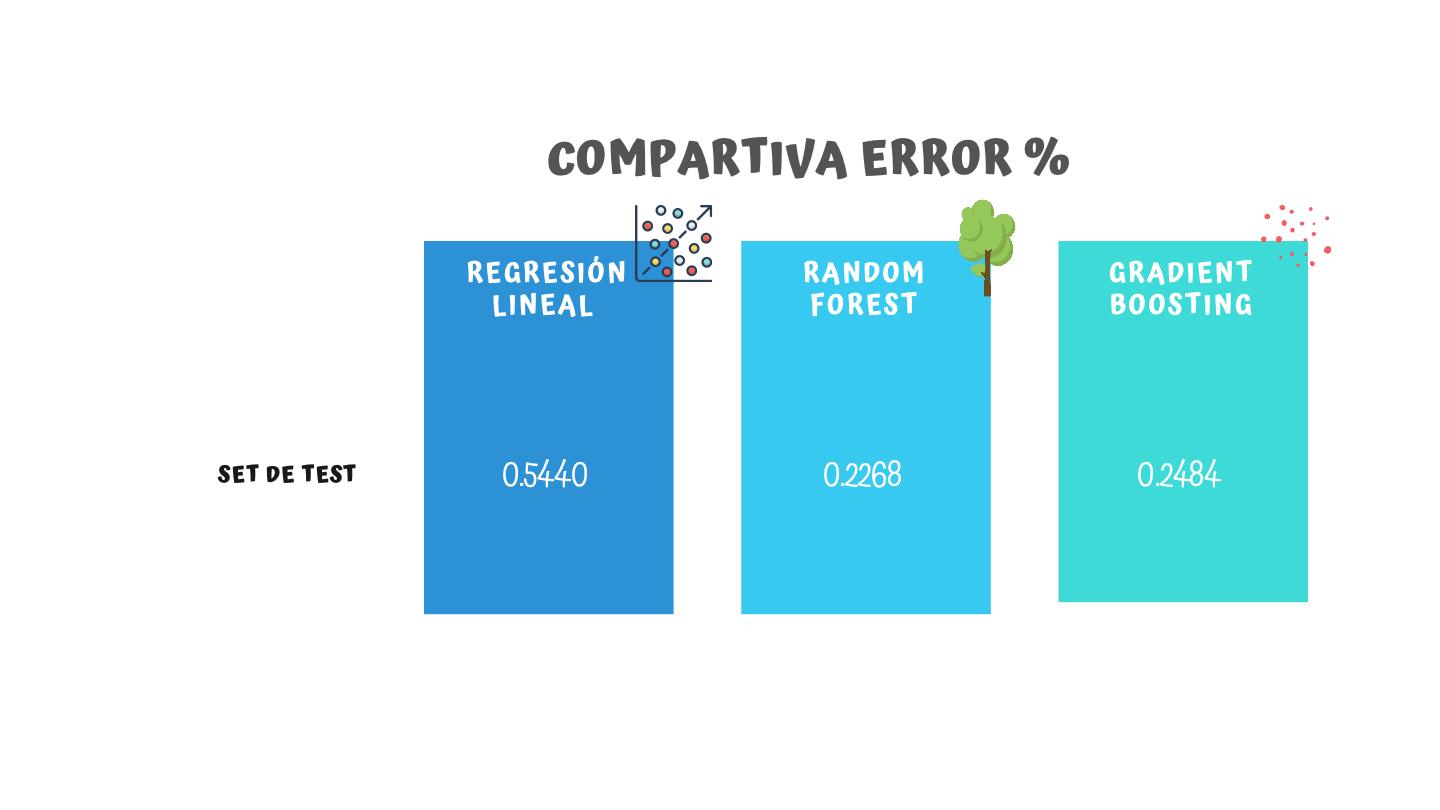
# Entrenamos y predecimos con dicho modelo

entrenar\_modelo\_y\_predecir(modelo\_GB)

# RESULTADOS y comparativas DE MODELOs

Las medidas que hemos obtenido en cada uno de los modelos han sido:





Decimos realizar este proyecto con el modelo **Random Forest** principalmente por lo que a resultados favorables hemos obtenido y por la sencillez a la hora analizar / limpiar los datos. Este modelo se considera como la “panacea” ya que es útil para regresión y clasificación, pero sobre todo sirve como una técnica para reducción de la dimensionalidad.

Los **Random Forest** consiste en, además de hacer Bootstrap y muestrear las variables de entrada en cada nodo del árbol. De esta forma, cada árbol se construiría utilizando sólo determinadas variables y no la totalidad de las variables de entrada. Este algoritmo sigue los siguientes pasos:

* Se toma un subconjunto de los datos.
* Se crea un árbol a partir del subconjunto, pero esta vez se dan pasos adicionales. Para cada nodo del árbol, se determina la mejor regla de partición utilizando aleatoriamente m variables de entrada de las n variables totales.
* Se repiten los pasos 1 y 2 múltiples veces.
* Se obtienen las predicciones de las observaciones OOB en cada árbol obtenido. La predicción final se obtiene por dos métodos:
  + Modelo de clasificación: utiliza el criterio del voto mayoritario.
  + Modelo de regresión: utiliza la media.

Algunas de las ventajas de este tipo de algoritmos son:

* Es uno de los algoritmos de aprendizaje más precisos.
* Puede manejar un gran número de variables aun cuando el número de observaciones es relativamente pequeño.
* Proporciona información sobre la importancia de cada una de las variables de entrada.
* Es un método eficaz para estimar datos incompletos.
* Robusto frente a valores atípicos.
* Ofrece información acerca de las interacciones entre las variables.

En lo que se refiere a precisión de la predicción, se puede considerar a Random Forest el vencedor general: se adapta muy bien a funciones diferentes, y llega a alcanzar predicciones de error cero. Aun así, Boosting demuestra muy buenos resultados de predicción, de hecho existen razones por las que Boosting puede ser más recomendable frente a modelos complejos o en los que se desconozca por completo la función asociada.

Por otro lado, Random Forest presenta picos de error inesperados y fenómenos de overfitting no previsibles, mientras que Boosting no, lo que hace pensar que el riesgo a priori es menor. En conclusión, Random Forest demuestra ser el método de predicción más potente, si bien su uso se recomienda dentro de condiciones conocidas o parcialmente conocidas. Boosting demuestra mayor homogeneidad de resultados y no presenta picos de error inesperados.

## Análisis de Residuos

Decidimos realizar un análisis de residuos con el modelo completo de la muestra para ver si podemos encontrar una explicación en las variables que tenemos menor predicción:

#VAMOS A REALIZAR EL ANALISIS DE LOS RESIDUOS

#Inicilizalizamos n al primer analisis de residuos que es del modelo de entrenamiento

n='ENTRENAMIENTO'

def plot\_tree\_residuals(model,X, y):

'''

Función para visualizar el residual plot de los modelos de arbol.

Es identica a la del testeo de la linealidad

'''

fitted\_vals = model.predict(X)

resids = y-fitted\_vals

fig, ax = plt.subplots(1,2)

sns.regplot(x=fitted\_vals,

y=y,

lowess=True,

ax=ax[0],

line\_kws={'color': 'red'})

ax[0].set\_title('Observado vs. Valores Predichos ' + n, fontsize=16)

ax[0].set(xlabel='Predichos', ylabel='Observado')

sns.regplot(x=fitted\_vals,

y=resids,

lowess=True,

ax=ax[1],

line\_kws={'color': 'red'})

ax[1].set\_title('Residuos vs. Valores Predichos ' + n, fontsize=16)

ax[1].set(xlabel='Predichos', ylabel='Residuos')

%matplotlib inline

%config InlineBackend.figure\_format ='retina'

sns.set\_style('darkgrid')

sns.mpl.rcParams['figure.figsize'] = (15.0, 9.0)

plot\_tree\_residuals(modelo\_RF, X\_train, y\_train)

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Como no conseguimos sacar una conclusión específica, creemos que lo mejor es realizar este análisis por zonas para tratar de identificar las zonas donde tenemos más residuos y encontrar una justificación o variabilidad que explique el modelo. Para ello realizamos un bucle que realice este análisis por cada zona previamente definida:

#Definimos una macro que recorra todos las zonas

for n in zonas:

if n!= 'ARGANZUELA':

print("Analizamos los residuos por zona: ", n)

gdf\_join\_predict=gdf\_join\_out[feature\_cols]

X\_zona=gdf\_join\_predict[gdf\_join\_predict['ZONA\_'+n]==1]

print('Las prediccion media es:' ,modelo\_RF.predict(X\_zona).mean())

y\_real=gdf\_join\_out[gdf\_join\_out['ZONA\_'+n]==1]

y\_real=y\_real['UNITPRICE']

print('El valor m2 es de:',y\_real.mean())

# Calidad predicción test

rmse\_zona = round(mean\_squared\_error(y\_real, modelo\_RF.predict(X\_zona), squared=False), 2)

print('RMSE en set de test :', rmse\_zona)

#Error en porcentaje

rmse\_zona\_porcentaje = rmse\_zona/y\_test.mean()

print('Error en % de test : ',rmse\_zona\_porcentaje)

plot\_tree\_residuals(modelo\_RF, X\_zona, y\_real)

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

# PRODUCTIVIZACIÓN

## Despliegue del modelo

Una vez tenemos ya el modelo definido, el siguiente paso es buscar la mejor forma de productivizar y llevar al negocio esta mejora.

Se va a reemplazar el modelo actual de asignación de precios por los nuevos modelos elaborados por el equipo de data analytics, realizados por la agrupación de barrios definida anteriormente.

Para estos nuevos modelos se va a introducir la mejora de que el cliente obtenga en real time el resultado de la valoración de la vivienda que quiere vender. Se generará una API que en tiempo real obtenga los datos introducidos por el cliente y ejecute el modelo el cual devolverá el resultado de la valoración.

Diagrama

Descripción generada automáticamente

Una vez generada la nueva API, se desplegará mediante el Canary Deployment, por el cual no vamos a sustituir el modelo antiguo por el nuevo de un día para otro, sino que ambos desarrollos van a convivir en un lapso de tiempo y en un grupo de barrios controlados con el fin de poder monitorear estadísticas y errores para luego reemplazar el código actual por los nuevos modelos.

Tabla, Icono

Descripción generada automáticamente

En la etapa 0 se ejecutará el modelo anterior, pasando a ejecutar los nuevos modelos en un grupo reducido de barrios en los que controlaremos los posibles errores a la hora de desplegar el modelo en producción.

Pasado un día en esta etapa, realizaremos el monitoreo de estos resultados y desplegamos la etapa 2 en la que incluiremos un nuevo listado de barrios en la que volveremos a monitorear durante un día para una vez concluido estos chequeos llegaremos a la etapa final donde se procederá a sustituir el código anterior por los nuevos modelos.

Estos nuevos anuncios se integrarán en las BBDD de cara a las futuras actualizaciones de estos modelos.

## Monitoreo del modelo

Una vez al mes se realizará el seguimiento de los modelos midiéndose estadísticas de estabilidad y desempeño.

Para medir la estabilidad utilizaremos el índice de estabilidad de la población (PSI) en el que tomaremos:

PSI < 0,1  --> Sin cambios significativos en la población

PSI < 0,2  --> Cambios moderados en la población

PSI > 0,2  --> Cambios significativos en la población

Cuando obtengamos que hay cambios significativos en la población será necesario el reentrenamiento de los modelos con una población más reciente.

Este indicador lo calcularemos comparando la población de desarrollo versus la población de los últimos tres meses, para tener una alerta temprana y versus los últimos seis meses para decidir si se entrena el modelo de nuevo.

Para medir el desempeño, utilizaremos la degradación del porcentaje del error.

Degradación < 0,1  --> Sin cambios significativos en la población

Degradación  > 0,1  --> Cambios significativos en la población

Cuando observemos cambios significativos en la población se procederá al reentrenamiento del modelo.

# conclusiones y próximos pasos

El desarrollo de un modelo con una precisión en la valoración del precio de la vivienda del 78%, hace que este trabajo cumpla con el objetivo de realización de un modelo similar a los presentes actualmente en producción en empresas de la competencia (precisión de valoración aproximada del 85%) y que, como hemos podido documentar, responden a una necesidad real y creciente en el sector inmobiliario.

Resultando conscientes del margen de mejora, nos encontramos convencidos de que a través de siguientes iteraciones podríamos continuar aumentando este porcentaje hasta asimilarlo en mayor medida.

A modo de próximos pasos que deberíamos de llevar a cabo si este proyecto siguiera su ciclo de vida en una empresa planteamos:

* Realizar una demo de la productivización del producto en la herramienta **Streamlit**. No hemos podido llevarlo a cabo
* Estudio de distorsión del modelo en zonas de mayor valor 🡪 análisis de residuos
* Peso de variables por zonas, no tienen nada que ver si las estudias en conjunto que por separado (ej: barrio, edificio rehabilitado, etc.)

De igual modo y como se ha planteado en el apartado introductorio, resultaría de gran interés contar con un dataset posterior al confinamiento de 2020 debido la crisis de la COVID-19 y poder observar si los pesos de las variables introducidas en el modelo han resultado modificados significativamente por la variación en los hábitos de consumo: búsqueda de espacios más amplios y mejor iluminados, tendencia de abandonar los centros urbanos con preferencia por las periferias, teletrabajo…

La propuesta de la introducción de la mejora de visualización de la tasación en tiempo real, también contaría como un siguiente paso, ya que como hemos podido observar a través de la documentación consultada, la Ciencia de Datos no solo se orienta a la optimización de procesos y procedimientos, sino que también se enfoca progresivamente y en gran medida a la mejora de la experiencia del usuario. Esta experiencia también está ampliamente marcada por hábitos de consumo fluctuantes y tendentes a la inmediatez y a la presentación de contenido más visual, aspectos que tendríamos en cuenta para las siguientes versiones de nuestro trabajo.